End User Documentation
Document Version: 1.0 – 2014-11

CUSTOMER

SAP InfiniteInsight® 7.0

Explorer - Sequence Coding User Guide

Table of Contents

Introduction to Application Scenarios 3

Scenario 1 ..... *4*
Scenario 2 ..... *4*

Introduction to Sample Files 5

Scenario 1: Segment Visitors to Understand Purchase Behavior Using File Counts 6

Step 1 - Selecting the Data *8*
Selecting a Data Source 8
Describing the Data 9
Selecting Events Data 13
Describing Events Data 14
Step 2 - Defining the Modeling Parameters *15*
Setting Sequence Coding Parameters 15
Selecting Sequence Coding Statistics 22
Checking the Transactions 24
Selecting Variables 26
Setting the Number of Clusters 27
Step 3 - Generating and Validating the Model *28*
Generating the Model 28
Validating the Model 29
Step 4 - Analyzing and Understanding the Model *31*
Segment Descriptions 31

Scenario 2: Predict End of Session Using Intermediate Sequences 32

Step 1 - Selecting the Data *33*
Step 2 - Defining the Modeling Parameters *33*
Setting Sequence Coding Parameters 33
Selecting Sequence Coding Statistics 33
Checking the Transactions 34
Selecting Variables 34
Step 3 - Generating and Validating the Model *34*
Generating the Model 34
Validating the Model 35
Step 4 - Analyzing and Understanding the Model *36*
Contributions by Variables 36
Significance of Categories 37

Introduction to Application Scenarios

In these scenarios, you are the Marketing Director of an E-commerce company and you want to increase the profitability of your Web site. You have the budget to launch a major marketing initiative, but you're not sure what kind of campaign would be the most effective. Due to market pressures, you only have the time and money to test a few campaigns before launching a major initiative. The two key metrics that are being used to measure the performance of the Web site are the "conversion rate" and "stickiness". The conversion rate of a site is the percentage of visits that result in a purchase. At this time, your Web site has a conversion rate of 4%, meaning that 4 out of every 100 visitors purchase at least one item. The stickiness of a Web site is a measure of the number of pages viewed by each visitor. The more pages a visitor views, the more likely they are to purchase something. Your Web site is averaging about 10 pages per visit.

In order to achieve rapid insight into the different groups of visitors to your Web site, you have decided to use *InfiniteInsight™ Modeler - Segmentation/Clustering* to group the population with respect to their buying behavior and site abandonment. The goal of the analysis is to get descriptions of the groups of visitors who tend to purchase items frequently, and the indicators that a session is about to end. You already know the following basic facts about your Web site:

- An average of 50,000 visitors come to the Web site each day
- For the 2000 sessions that result in a purchase each day, the average amount spent is $181
- The average profit margin for the Web site is 5%, so each purchase results in an average profit of $9.05, resulting in $18,100 of profit per day.
- There are four main entry points for the site – The home page, the members home page, the sweepstakes page, and the specials page.
- The checkout process has five steps, all with the word "order" in the file name.
- Your site does not use "cookies" or require a login for your members, so each session is effectively anonymous unless a purchase is made.

The information that is available for analysis consists of the Web logs. Your DBA has pulled out a list of the sessions from a single day of traffic, along with a flag indicating if the session resulted in a purchase (the existence of "order5.tmpl" in a session indicates a purchase). Along with the list of sessions, the parsed log from the day is also available. Since the information from the Web log is not aggregated for analysis, you will need to use the *InfiniteInsight™ Explorer - Sequence Coding* prior to running the *InfiniteInsight™ Modeler - Segmentation/Clustering* or *InfiniteInsight™ Modeler - Regression/Classification*.

In This Chapter

Scenario 1 4
Scenario 2 4

Scenario 1

You will start by using Sequence Coding to create counts of each Web page that was viewed by each visitor, followed by a targeted segmentation with “purchase” as the target. This will give you a simple description of the different groups browsing your Web site, and the different conversion rates for each group.

Scenario 2

In this scenario you want to predict when a visitor is going to leave your Web site. Your idea is to offer a $5 coupon to visitors who are likely to leave in the hope of increasing the site stickiness. To achieve that, you will create a Sequence Coding model using **intermediates sequences** with the **FirstLast** option for the pages viewed. The intermediate sequence option will automatically create an appropriate target variable for determining which behaviors indicate the end of a session.

Introduction to Sample Files

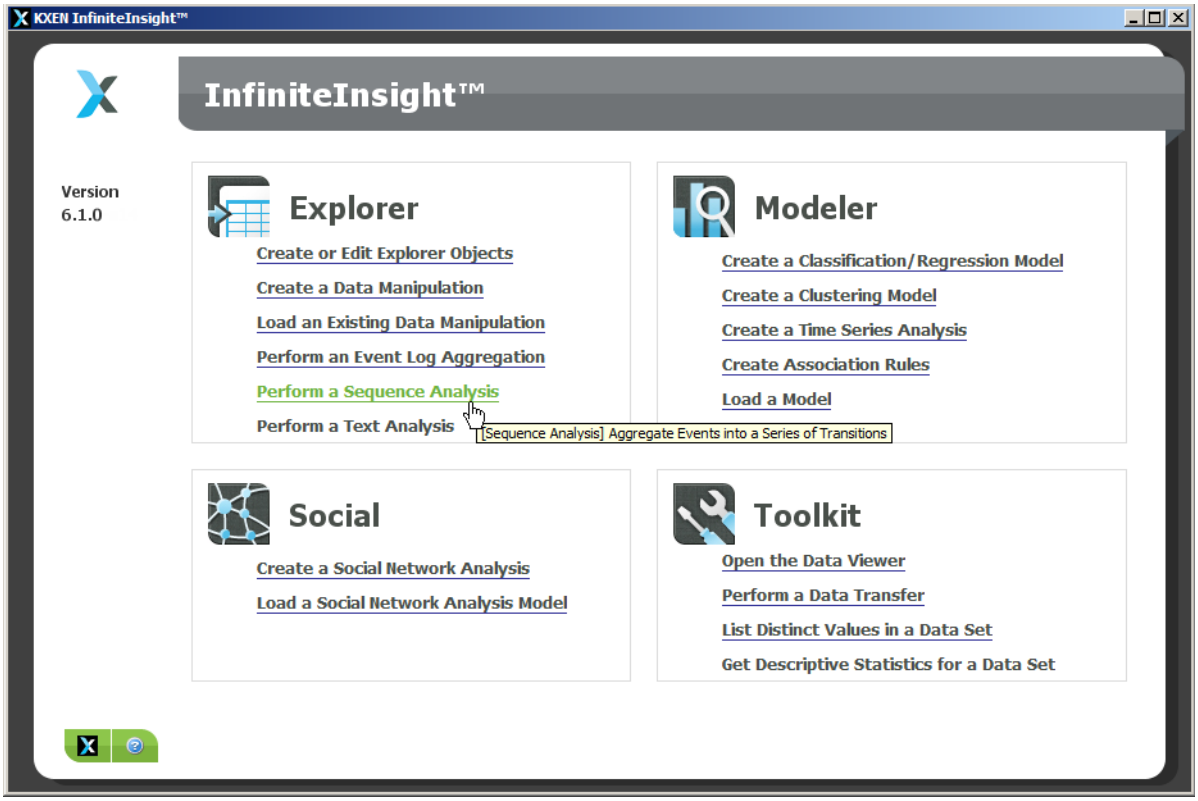
This data set contains a single day of Web traffic from an E-commerce site in December 1999. The site content was served by a Broadvision server, but no "cookies" or login was required, making the sessions effectively anonymous.

File	Description
session_purchase.csv	list of sessions and binary purchase target (50581 rows)
session_purchase_desc.csv	description for session_purchase.csv
file_view.csv	log of files requested from Broadvision server (532860 rows)
file_view_desc.csv	description for file_view.csv
session_purchase_skip.csv	variable skip list for Scenario 1 (these are the variables where the value would not be known until the session had ended)
session_continue_skip.csv	variable skip list for Scenario 2

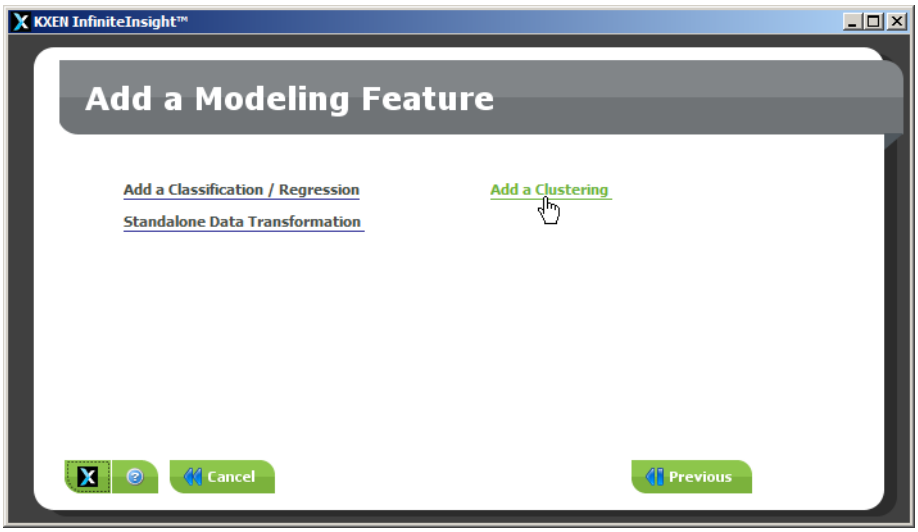
These sample files can be downloaded on *InfiniteInsight™* **Sample Files Download Center** (**http://www.kxen.com/sample_data/**).

Scenario 1: Segment Visitors to Understand Purchase Behavior Using File Counts

1 In *InfiniteInsight™* main menu, select the option **Perform a Sequence Analysis** in the **Explorer** section.



2 The screen **Add a Modeling Feature** is displayed.



3 Click on the option **Add a Clustering**.

Note - When building a model you can either simply analyze the sequences or add extra transformations such as a Classification/Regression (*InfiniteInsight™ Modeler - Regression/Classification*) or a Clustering/Segmentation (*InfiniteInsight™ Modeler - Segmentation/Clustering*)

IN THIS CHAPTER

Step 1 - Selecting the Data 8
Step 2 - Defining the Modeling Parameters 15
Step 3 - Generating and Validating the Model 28
Step 4 - Analyzing and Understanding the Model 31

Step 1 - Selecting the Data

IN THIS CHAPTER

Selecting a Data Source 8
Describing the Data 9
Selecting Events Data 13
Describing Events Data 14

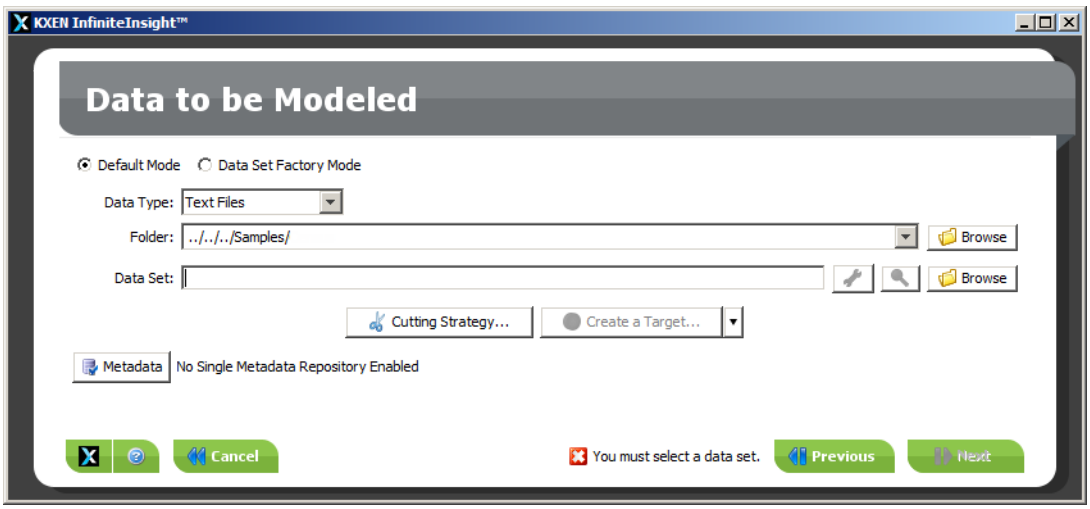
Selecting a Data Source

For this Scenario

The file **session_purchase.csv** contains a list of session IDs and whether each session has led to a purchase or not. This will be referred to as the Reference data set for Sequence Coding. A Sequence Coding Reference data set must have a single variable unique primary key. If the primary key is non-unique or spread out over several variables, Sequence Coding will not function properly.

☑ To Select a Data Source

1 On the screen **Data to be Modeled**, select the data source format to be used (**Text files**, **ODBC**, ...).



2 Use the **Browse** button on the right of the **Folder** field to select the folder where you have saved the sample files.

3 Click the **Browse** button next to the **Estimation** field and select the file **session_purchase.csv**. The name of the file will appear in the **Estimation** field.

4 Click the **Next** button.

Describing the Data

Why Describe the Data Selected?

In order for the *InfiniteInsight™* features to interpret and analyze your data, the data must be described. To put it another way, the description file must specify the nature of each variable, determining their:

- **Storage** format: number (**number**), character string (**string**), date and time (**datetime**) or date (**date**).

Notes -

When a variable is declared as **date** or **datetime**, the KXEN Date Coder feature (*KDC*) automatically extracts date information from this variable such as the day of the month, the year, the quarter and so on. Additionnal variables containing this information are created during the model generation and are used as input variables for the model.

KDC is disabled for Time Series.

- **Type: continuous**, **nominal**, **ordinal** or **textual**.

For more information about data description, see **Types of Variables** and **Storage Formats** in the **Introductory Guide to** *InfiniteInsight™*.

How to Describe Selected Variables

To describe your data, you can:

- Either **use an existing description file**, that is, taken from your information system or saved from a previous use of *InfiniteInsight™* features,
- Or **create a description file using the Analyze option**, available to you in *InfiniteInsight™*. In this case, **it is important that you validate the description file obtained**. You can save this file for later re-use. If you name the description file **KxDoc_<SourceFileName>**, it will be automatically loaded when clicking the **Analyze** button.

Important - The description file obtained using the **Analyze** option results from the analysis of the first 100 lines of the initial data file. In order to avoid all bias, we encourage you to mix up your data set before performing this analysis.

Each variable is described by the fields detailed in the following table:

The Field...	Gives information on...
Name	the variable name (which cannot be modified)
Storage	the type of values stored in this variable: ▪ **Number**: the variable contains only "computable" numbers (be careful a telephone number, or an account number should not be considered numbers) ▪ **String**: the variable contains character strings ▪ **Datetime**: the variable contains date and time stamps ▪ **Date**: the variable contains dates

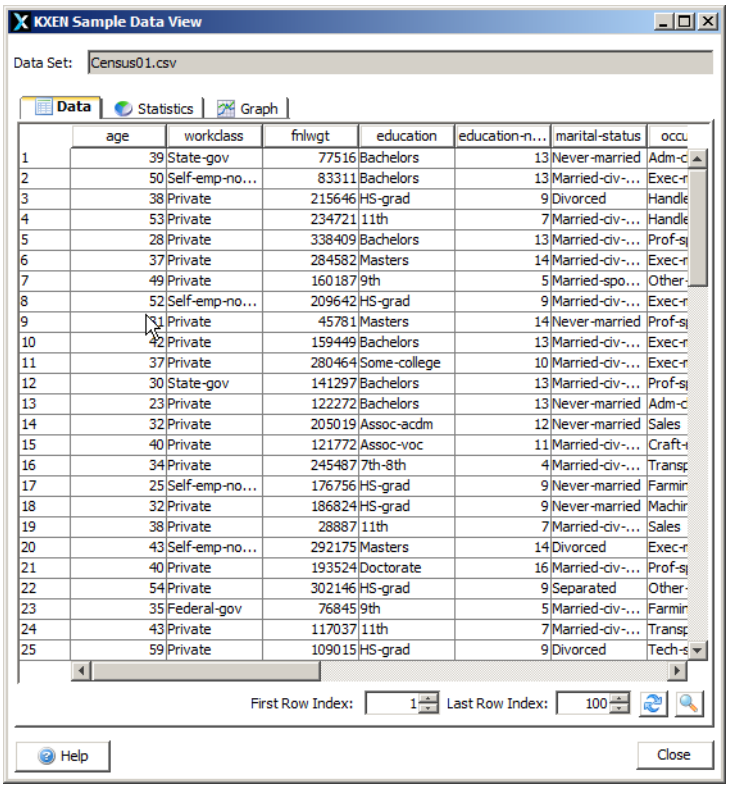
The Field...	Gives information on...
Value	the value type of the variable: ▪ **Continuous**: a numeric variable from which mean, variance, etc. can be computed ▪ **Nominal**: categorical variable which is the only possible value for a string ▪ **Ordinal**: discrete numeric variable where the relative order is important ▪ **Textual**: textual variable containing phrases, sentences or complete texts Warning - When creating a text coding model, if there is not at least one textual variable, you will not be able to go to the next panel.
Key	whether this variable is the key variable or identifier for the record: ▪ **0** the variable is not an identifier; ▪ **1** primary identifier; ▪ **2** secondary identifier...
Order	whether this variable represents a natural order. (0: the variable does not represent a natural order; 1:the variable represents a natural order). If the value is set at 1, the variable is used in SQL expressions in an "`order by`" condition. There must be at least one variable set as Order in the Event data source. Warning - If the data source is a file and the variable stated as a natural order is not actually ordered, an error message will be displayed before model checking or model generation.
Missing	the string used in the data description file to represent missing values (e.g. "999" or "#Empty" - without the quotes)
Group	the name of the group to which the variable belongs. Variables of a same group convey a same information and thus are not crossed when the model has an order of complexity over 1 . This parameter will be usable in future version.
Description	an additional description label for the variable
Structure	this option allows you to define your own variable structure, which means to define the variables categories grouping.

Viewing the Data

To help you validate the description when using the **Analyze** option, you can display the first hundred lines of your data set.

☑ To View the Data

1 Click the button **View Data**. A new window opens displaying the data set top lines:



	age	workclass	fnlwgt	education	education-n...	marital-status	occu
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-c
2	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	Exec-r
3	38	Private	215646	HS-grad	9	Divorced	Handle
4	53	Private	234721	11th	7	Married-civ-...	Handle
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-sp
6	37	Private	284582	Masters	14	Married-civ-...	Exec-r
7	49	Private	160187	9th	5	Married-spo...	Other-
8	52	Self-emp-no...	209642	HS-grad	9	Married-civ-...	Exec-r
9	31	Private	45781	Masters	14	Never-married	Prof-sp
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-r
11	37	Private	280464	Some-college	10	Married-civ-...	Exec-r
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-sp
13	23	Private	122272	Bachelors	13	Never-married	Adm-c
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales
15	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-r
16	34	Private	245487	7th-8th	4	Married-civ-...	Transp
17	25	Self-emp-no...	176756	HS-grad	9	Never-married	Farmin
18	32	Private	186824	HS-grad	9	Never-married	Machin
19	38	Private	28887	11th	7	Married-civ-...	Sales
20	43	Self-emp-no...	292175	Masters	14	Divorced	Exec-r
21	40	Private	193524	Doctorate	16	Married-civ-...	Prof-sp
22	54	Private	302146	HS-grad	9	Separated	Other-
23	35	Federal-gov	76845	9th	5	Married-civ-...	Farmin
24	43	Private	117037	11th	7	Married-civ-...	Transp
25	59	Private	109015	HS-grad	9	Divorced	Tech-s

2 In the field **First Row Index**, enter the number of the first row you want to display.

3 In the field **Last Row Index**, enter the number of the last row you want to display.

4 Click the **Refresh** button to see the selected rows.

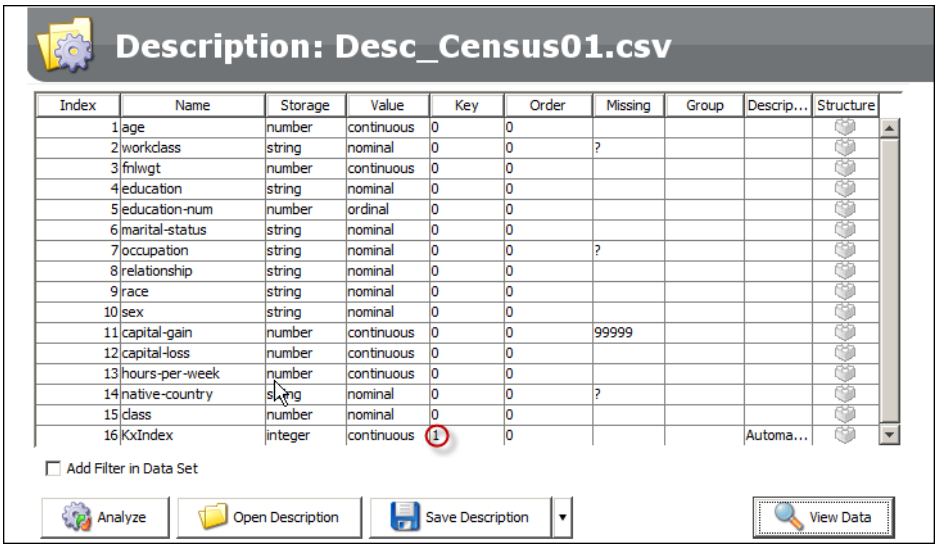
A Comment about Database Keys

For Sequence Coding to be able to join the Reference and Transaction data sets, the Reference data set to be analyzed must contain a single variable that serves as a unique key variable.

To Specify that a Variable is a Key

1 In the **Key** column, click the box corresponding to the row of the key variable.

2 Type in the value "**1**" to define this as a key variable.



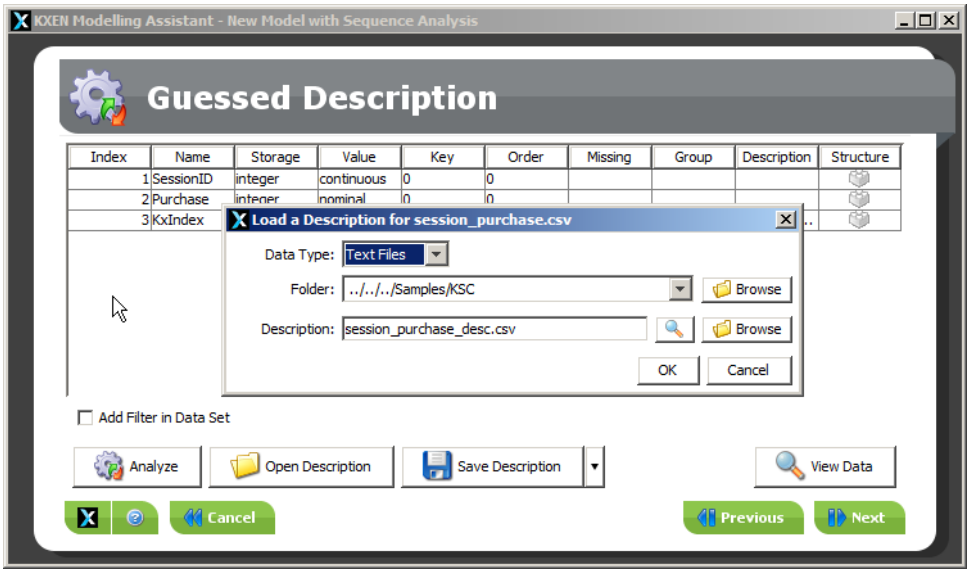
Index	Name	Storage	Value	Key	Order	Missing	Group	Descrip...	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-num	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-week	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automa...	

For this Scenario

Use the file **session_purchase_desc.csv** as the description file.

To Describe the Data

1 On the screen **Data Description**, click the button **Open Description**. The following window opens:



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	SessionID	integer	continuous	0	0				
2	Purchase	integer	nominal	0	0				
3	KxIndex								

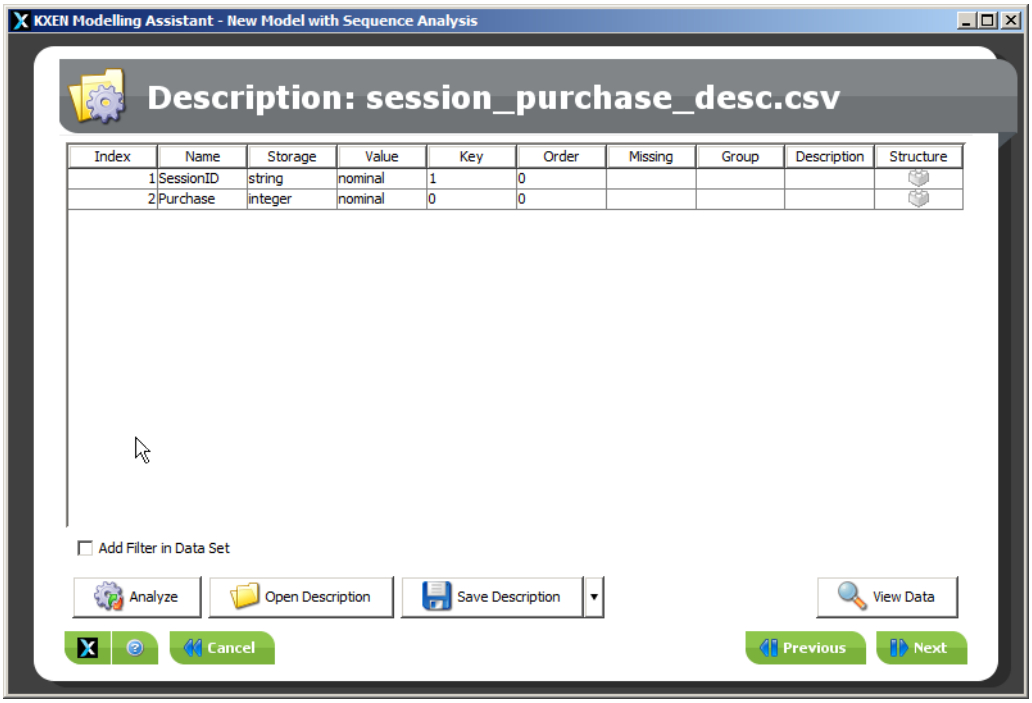
2 In the window **Load a Description**, select the type of your description file.

3 In the **Folder** field, select the folder where the description file is located with the **Browse** button.

Note - The folder selected by default is the same as the one you selected on the screen **Data to be Modeled**.

4 In the **Description** field, select the file containing the data set description with the **Browse** button.

5 Click the **OK** button. The window **Load a Description** closes and the description is displayed on the screen **Data Description**.



6 Click the **Next** button.

Selecting Events Data

The screen **Events Data** lets you specify the data source to be used as the Transaction data set.

For this Scenario

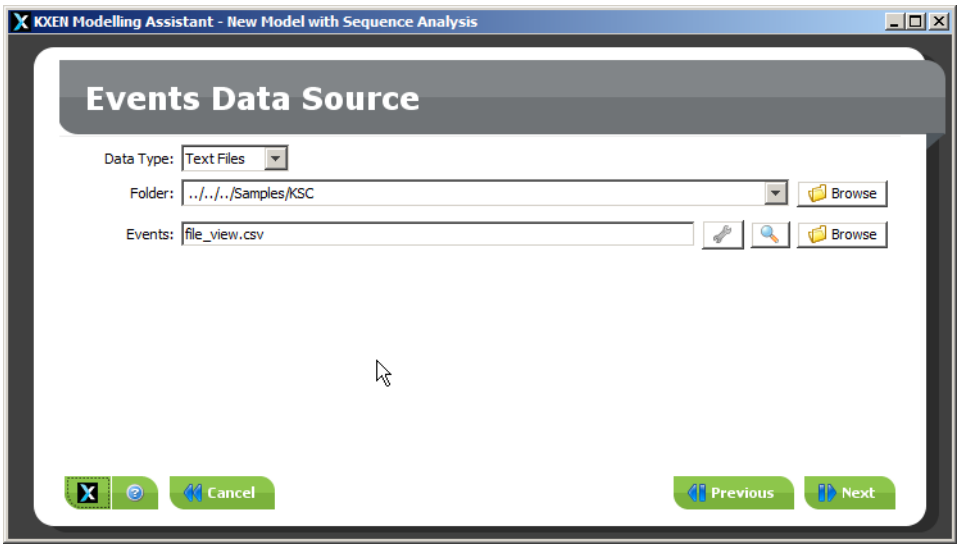
- The **Folder** field should already be filled in with the name of the data source that you specified on the **Data to be Modeled** screen.
- Select the file **file_view.csv**.

☑ To Select Events Data

1 Select the format of your data source (**Text Files**, **ODBC**, ...).

2 In the **Folder** field, specify the folder where your data source is stored.

3 In the **Events** field, specify the name of your data source.



4 Click the **Next** button.

Describing Events Data

The screen **Events Data Description** lets you describe your Transaction data, offering you the same options as the screen **Data Description**.

For Sequence Coding to function properly, there must be a variable in the Transaction data set that is the same as the primary key declared for the Reference data set, referred to as a “Join Column”. The name of the variable can be different, but the storage and value must be the same. The values of this variable need not be unique, since each Reference key can have 0, 1, or several associated transactions.

In addition to a suitable join column, the Transaction data set must have at least one datetime variable. The datetime variable will be used by Sequence Coding to order the transactions.

One of the datetime variables must absolutely be **ordered** and declared as such by setting to **1** the **Order** column for this variable in the description file.

When the data source comes from a database, *InfiniteInsight™* uses a query with an `order by` on the variable set as **Order** to retrieve the data. But when the data source is a file (.txt, .csv, ...), *InfiniteInsight™* verifies if the variable set as **Order** is actually ordered in the file, if not an error message is displayed.

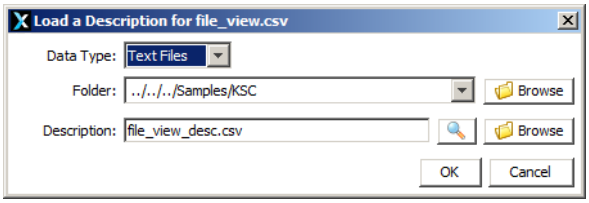
For detailed procedures on how to set parameters on this screen, see **Describing the Data** (on page 9).

For this Scenario

Use the description file **file_view_desc.csv**.

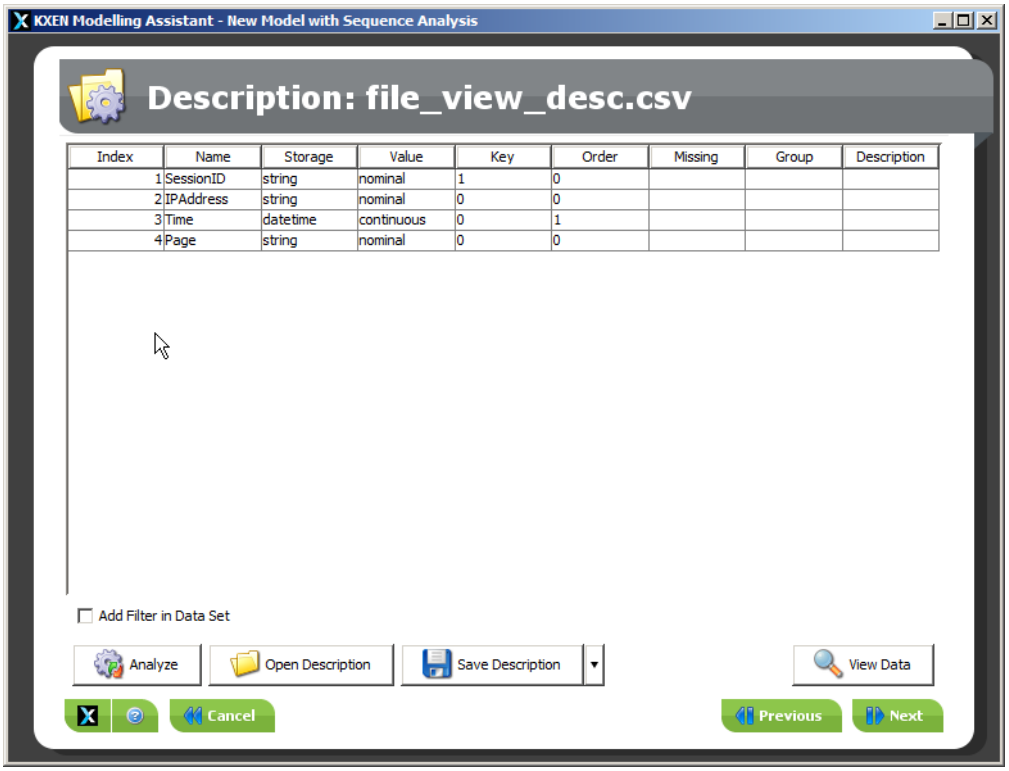
☑ To Describe Events Data

1 On the screen **Event Data Description**, click the button **Open Description**. The following window opens:



2 In the window **Load a Description**, select the file **file_view_desc.csv**.

3 Click the **OK** button. The window **Load a Description** closes and the description is displayed on the screen **Event Data Description**.



Note that the **Order** column is set at **1** for the **Time** variable, thus indicating that this variable is used as a natural order.

4 Click the **Next** button.

Step 2 - Defining the Modeling Parameters

Setting Sequence Coding Parameters

The screen **Sequence Analysis Parameters Settings** enables you to set some Sequence Coding parameters by performing the following tasks:

- Join your reference data with your transaction data
- Calculate the intermediate sequences
- Filter your events by period

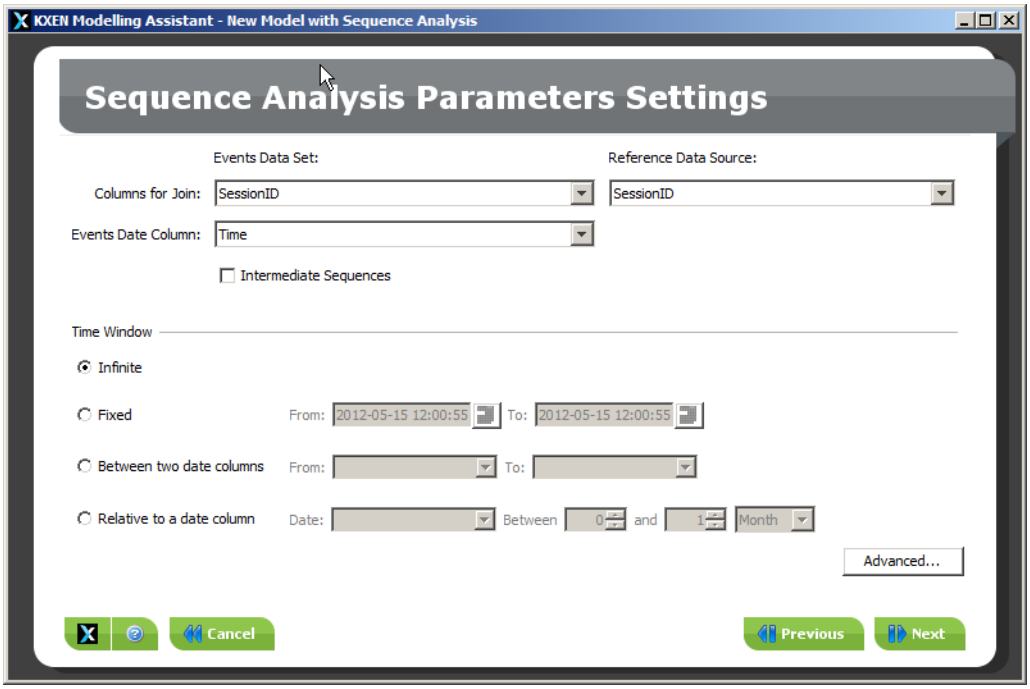
For this Scenario

- Select the `SessionID` column as the join column for both the log and reference data sets.
- Select `Time` as the **Log Date Column**.
- In the advanced parameters, keep **75%** of the hits.
- Select `Infinite` as the **Time Window**.

To Set the Parameters

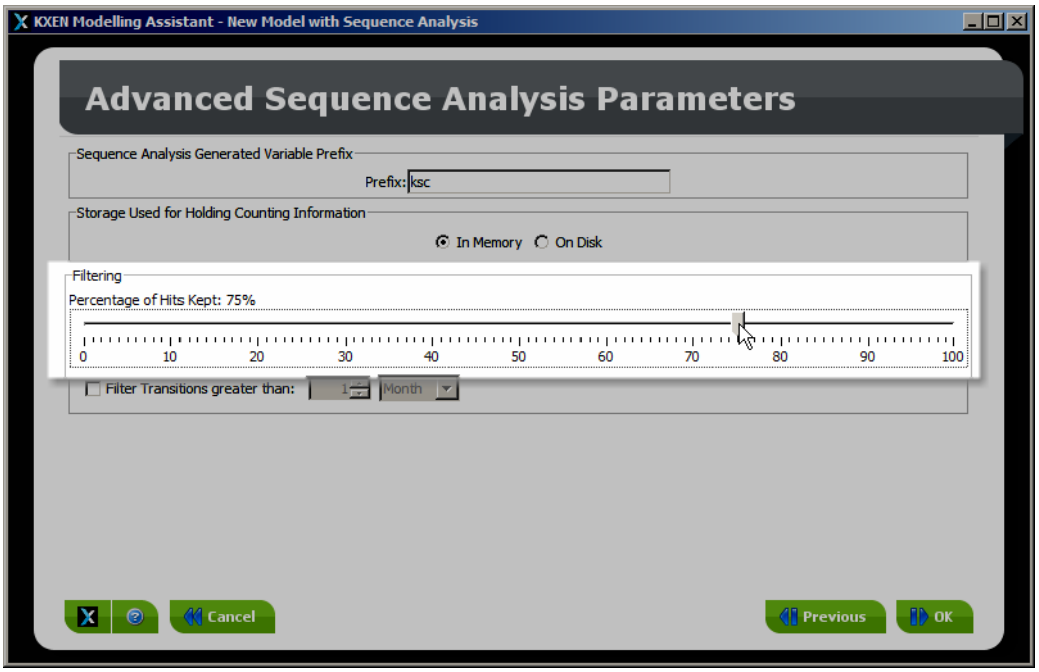
1 On the screen **Sequence Analysis Parameters Settings**, select the join column for both the log and reference data sets.

2 Select the **Log Date Column**.



3 Click the **Advanced** button to set the advanced parameters.

4 In the **Advanced** panel, slide the filter to **75%**.



Understanding InfiniteInsight™ Explorer - Sequence Coding Parameters

Joining Your Data

To aggregate the reference data with the events data, you have to join both tables and indicate which column of each table corresponds to the reference ID.

In the fields **Columns for Join**, select the variables corresponding to the customer ID in both data sets. The information contained in both selected variables must be the same.

In the field **Log Date Column**, select the variable corresponding to the date and/or time of the log data.

Calculating the Intermediate Sequences

The mode **Intermediate Sequences** provides you with additional information about the transitions and sequences existing in your data sets:

- order of the steps
- details of the steps
- continuity of the session for each step

Filtering the Events by Time Window

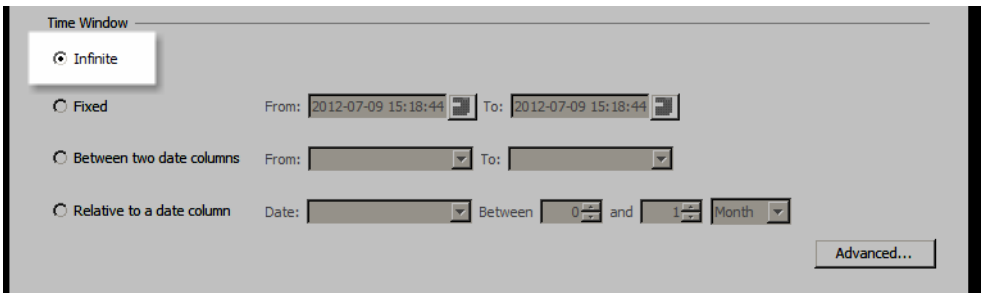
The section **Time Window** allows you to filter the events on which the model will be built by setting a period defined either by fixed dates or by values existing in the data set. The following options are available to filter the events data set:

Option	Description
Infinite	No time window is defined: all the events will be used.
Fixed	Only the events for which the **Log Date Column** value is between the two selected dates will be used.
Between two date columns	Only the events for which the **Log Date Column** value is between the values of the two selected date columns will be used. For example, you can select the date columns corresponding to the beginning and the end of a trial period, dates that can be different for each customer.
Relative to a date column	Only the events for which the **Log Date Column** value fits in the range defined with respect to the selected date column will be used. For example, you can use the purchase date of a credit card as the reference and select all events that occurred in the three months leading to this date.

WARNING - Be careful when choosing a period, the selected period must contain events existing in the data set, or else you will obtain aberrant results for your model (negative KI, KR equal to 1, ...).

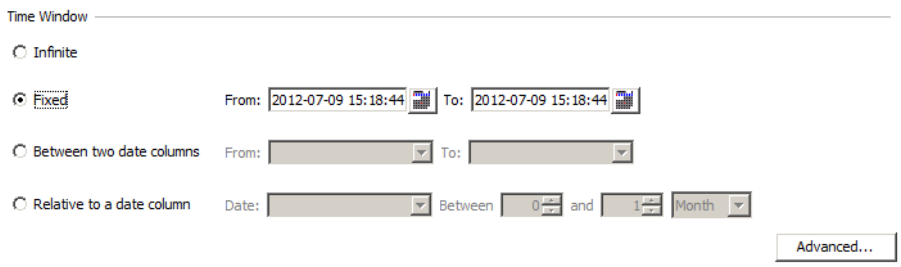
☑ To Use All the Events

Keep the **Infinite** option.



☑ To Use Only the Events Occurring in a Fixed Time Window

1 Check the **Fixed** option.

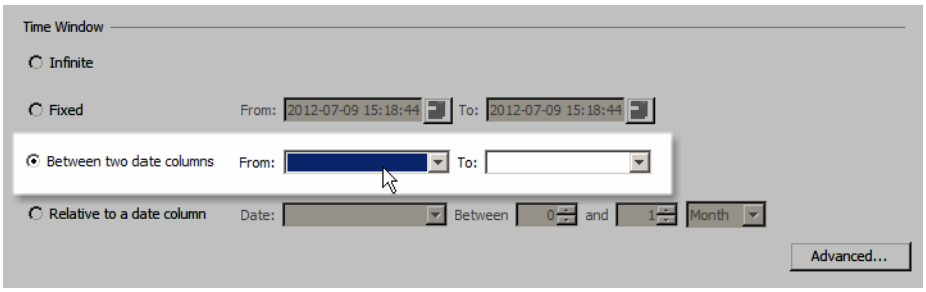


2 In the **From** field, select the date before which no events should be used.

3 In the **To** field, select the date after which no events should be used.

☑ To Use Only the Events Occurring Between Two Date Columns

1 Check the option **Between two date columns**.

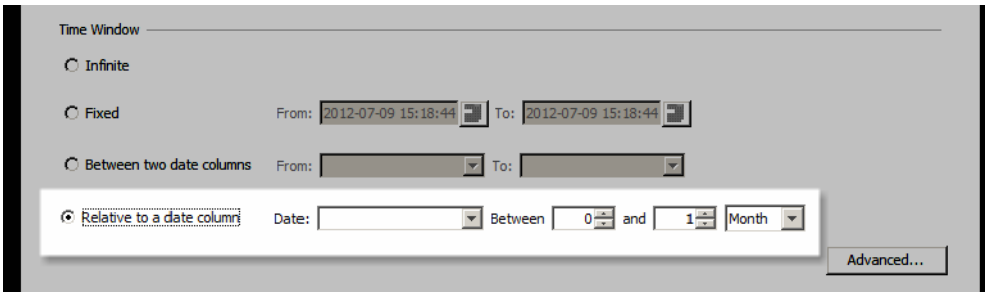


2 In the **From** field, select the date column containing the date before which no events should be used.

3 In the **To** field, select the date column containing the date after which no events should be used.

☑ To Use Only the Events Occurring in a Range Relative to a Date Column

1 Check the option **Relative to a date column**.



2 In the **Date** list, select the column that contains the date to use as a reference for the time window.

3 In the **Between** field, enter the number of units that will indicate the start of the time window. The following table sums up the values you can use to define the beginning of the time window.

Value	Significance
negative integer	the time window begins before the reference date
0	the time window begins at the reference date
positive integer	the time window begins after the reference date

4 In the **and** field, enter the number of units that will indicate the end of the time window.

5 In the last drop-down list, enter the unit to be used to define the time window.

For example, if you have set the parameters **Date** `CardPurchaseDate` **Between** `-3` **and** `0` `Month`, only events occurring in the three months leading to the date of purchase will be kept for each customer.

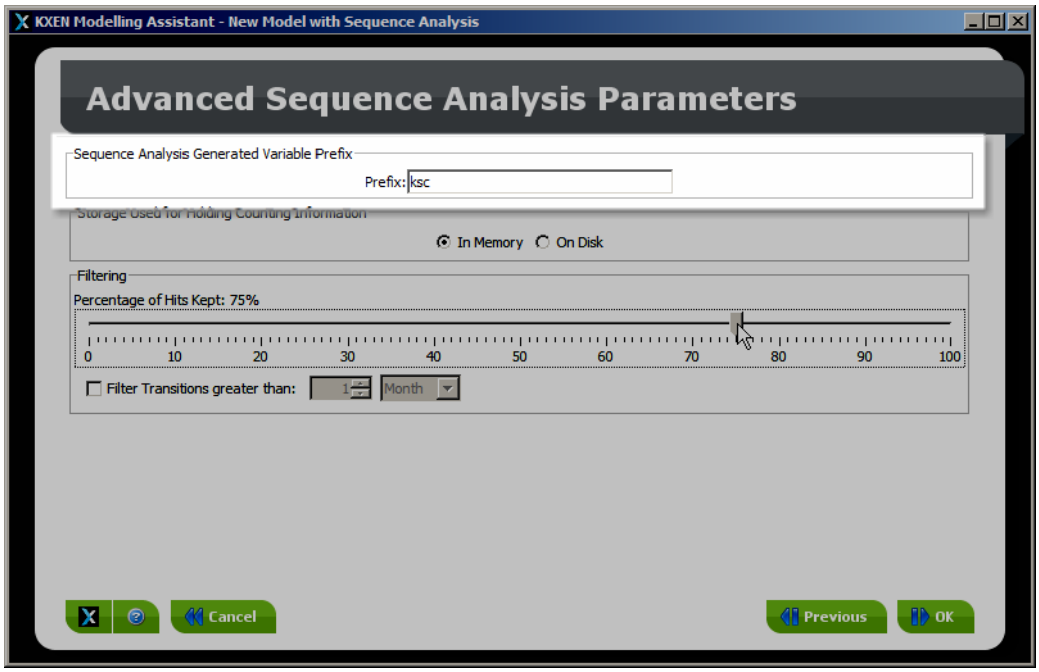
Understanding Advanced Parameters

The advanced parameters allows you to configure the following elements:

- the prefix to be added to Sequence Coding generated variables,
- the location where the temporary files generated by the modeling are stored,
- the amount of information that will be kept for the modeling.

Sequence Coding Generated Variable Prefix

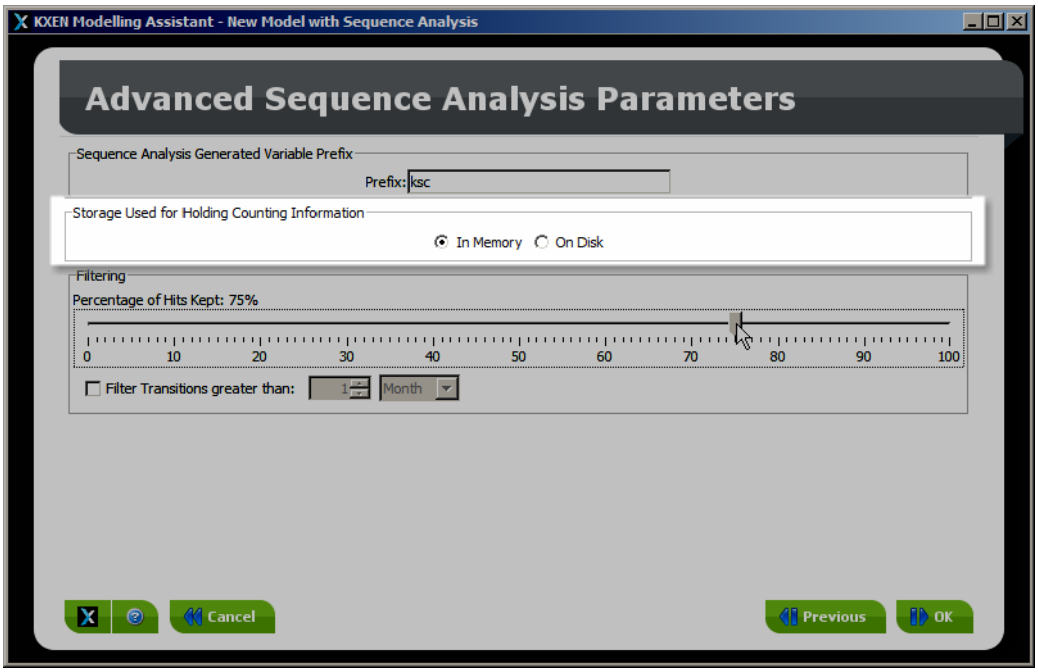
You can define a specific prefix that will be used to identify variables created by *InfiniteInsight™ Explorer*. By default, this prefix is set to `ksc`.



Storage Type

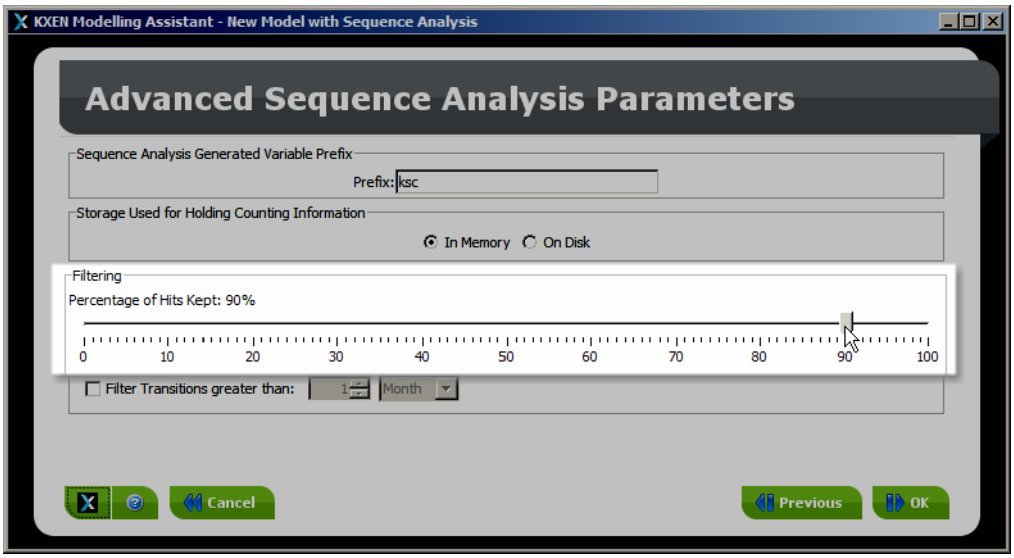
When creating a model, Sequence Coding generates large quantities of temporary columns, you can select whether the data generated will be stored in a memory space or on a disk.

The option **In memory** is selected by default.



Filtering the Events

The **Filtering** option allows you to group rare categories into a single category labeled KxOther. It is very common for transaction logs to have many infrequently occurring categories that by themselves will not make reliable predictors. A predictive benefit can often be achieved by combining these rare categories into a single group. The Filtering slide allows you to select the categories to keep as separate columns based on percentage of the overall transaction log. The categories corresponding to the remaining percentage of transactions are grouped in the **KxOther** column, which is automatically generated by *InfiniteInsight™ Explorer - Sequence Coding*.

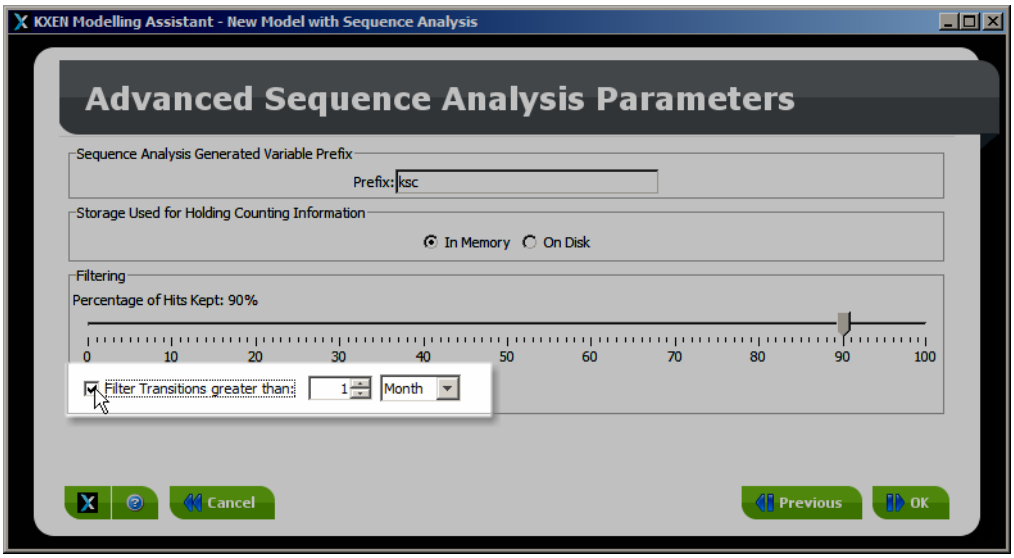


For example, if you set the **Filtering** slider at 90%, it means that the total number of transactions when adding all the categories assigned to separate columns must not exceed 90% of the total number of transactions. The categories that make up the remaining 10% of the transactions will be grouped under **KxOther**.

You can also define a **threshold** so that transitions which duration between two events is higher than the defined threshold will be ignored in the transition count.

☑ To Set a Threshold

1 Check the box **Filter Transitions greater than**.



2 In the number field, enter the number of units defining the threshold.

3 In the drop-down list, select the unit to be used to define the threshold.

For this Scenario

For the sample data, each row of the transaction log represents an HTML file requested by the visitor's browser. There are 10184 different files that are requested during the day. However, by positioning the Filtering slide at 75%, only 99 files are retained for separate count columns, and the rows with the remaining 10085 files are grouped into the KxOther count. This means that the 99 most common files make up 75% of the log and the remaining 10085 files make up only 25% of the log.

Selecting Sequence Coding Statistics

The screen **Sequence Analysis Variables Selection for Functions** lets you specify the type of statistics you want to calculate on transaction or event data.

For this Scenario

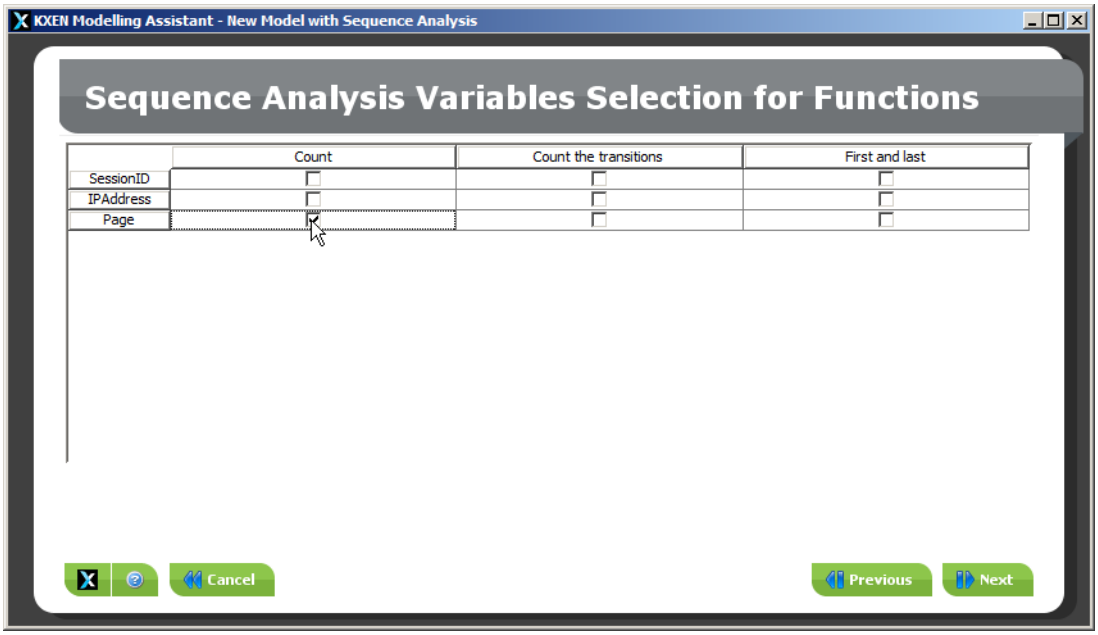
You decide to calculate for each session which pages have been visited on the web site. That way, you should be able to determine and understand which pages led the visitors to make a purchase.

You must use the following settings:

- For the variable **Page**, select the function **Count**, which will create a state column for each page visited.

☑ To Select Sequence Coding Statistics

1 The **Sequence Analysis Variables Selection for Functions** screen lists all the variables for which statistics can be calculated. For each variable listed, select the functions to use. You can choose among the three functions **Count**, **CountTransition** and **FirstLast**.



2 Click the **Next** button.

Operations Definition

Several standard Sequence Coding columns are created for each reference ID. For reference Ids that have no transactions associated with them, the standard Sequence Coding columns will have null values.

KSC_Start_Date: The timestamp of the first transaction in the log for each reference ID.

KSC_End_Date: The timestamp of the last transaction in the log for each reference ID.

KSC_TotalTime: The seconds between the **KSC_Start_Date** and **KSC_End_Date**.

KSC_Number_Events: The number of transactions in the log associated with each reference ID.

In addition to the standard *InfiniteInsight™ Explorer - Sequence Coding* columns, three types of operations are available:

- **Count**,
- **Count the transitions**,
- **First and last**.

Count

When you select the **Count** option, Sequence Coding creates a new column for each value of the inserted variables.

Count encodes the sequences using one column per valid category in the specified nominal column. Each valid category is referred to as a "state". Categories that are seen only once for the transactions associated with the reference id present in the Estimation data set are discarded.

CountTransition

When you select the **CountTransition** option, Sequence Coding creates a new column for each transition of categories in the selected data set.

CountTransition encodes the sequences using one column per valid pair wise category transition in the specified column. Each valid category transition is referred to as a "state transition". State transitions that are seen only once for the transactions associated with the reference id present in the **Estimation** data set are discarded. A separate **KxOther** column will be created for rare transitions, using the threshold set by the Filter slider bar in the same way a **KxOther** column is created for the counts.

FirstLast

The **FirstLast** option creates two columns, the categories of the selected variable from the first and last transactions in the log for each reference ID, called **FirstState** and **LastState** respectively. The **FirstState** and **LastState** columns are created automatically when either the **Count** or **CountTransition** options are selected.

Checking the Transactions

At this stage, *InfiniteInsight™* analyses the data sets and creates a number of new variables, or columns. Depending on which operations you chose during the previous step, Sequence Coding creates:

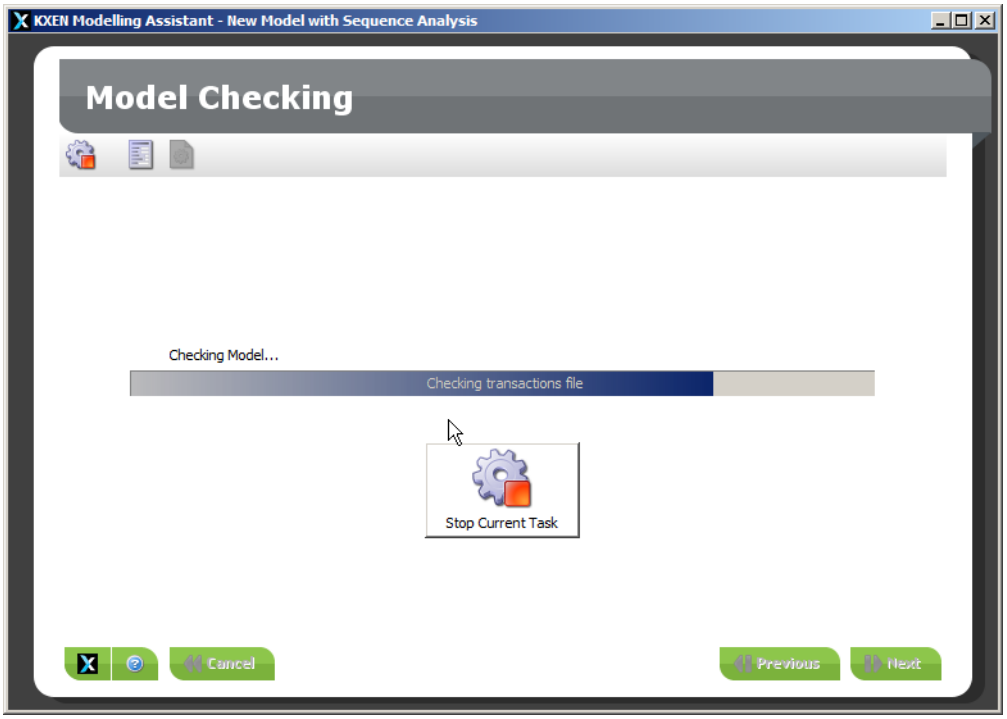
- four standard columns - **ksc_Start_Date**, **ksc_End_Date**, **ksc_TotalTime**, and **ksc_Number_Events**,
- one column for each state (if you have selected **Count**),
- one column for each transition (if you have selected **CountTransitions**).
- Two columns, **FirstState** and **FinalState** (if you have selected, **Count**, **CountTransitions**, or **FirstLast**).
- Six columns, **LastStepNumber**, **Last_date-time**, **Last_duration**, **Session_Continue**, **LastState**, and **NextState** (if you have selected **Intermediate Sequences**).

For this Scenario

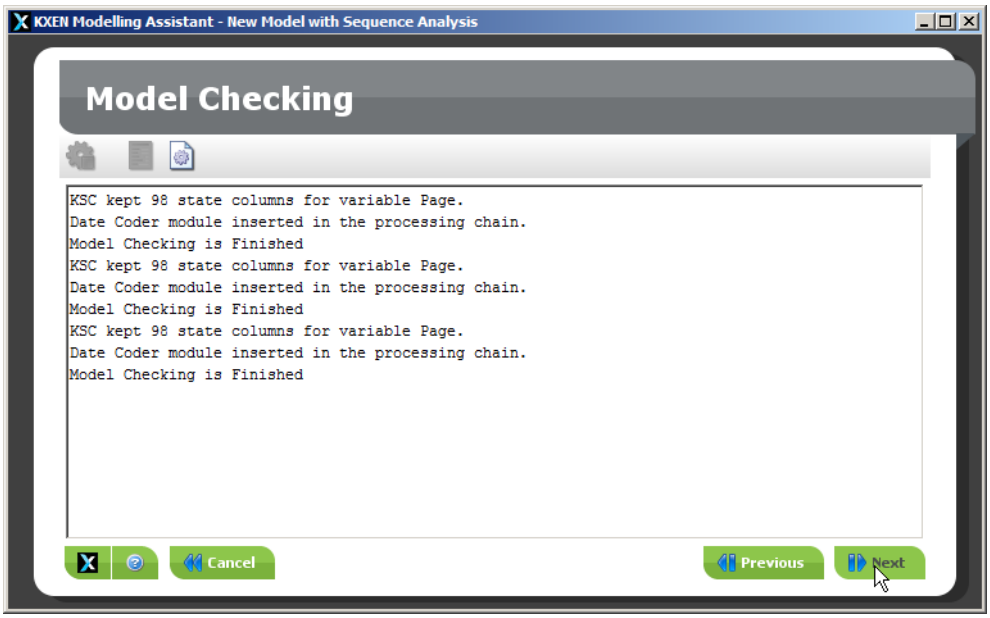
After the transactions are checked, Sequence Coding should have kept **99** state columns for the **Page** variable, plus the four standard columns and the **FirstState** and **LastState** columns.

☑ To Check the Transactions

1 During the model checking a progress bar is displayed.



2 When the process is over, click the button (**Show Detailed Log**). The number of columns created by Sequence Coding is indicated.


KXEN Modelling Assistant - New Model with Sequence Analysis
Model Checking
KSC kept 98 state columns for variable Page.
Date Coder module inserted in the processing chain.
Model Checking is Finished
KSC kept 98 state columns for variable Page.
Date Coder module inserted in the processing chain.
Model Checking is Finished
KSC kept 98 state columns for variable Page.
Date Coder module inserted in the processing chain.
Model Checking is Finished
Cancel
Previous
Next

3 Click the **Next** button.

Selecting Variables

Once the reference data set, the events data set and their descriptions have been entered, you must select different variables :

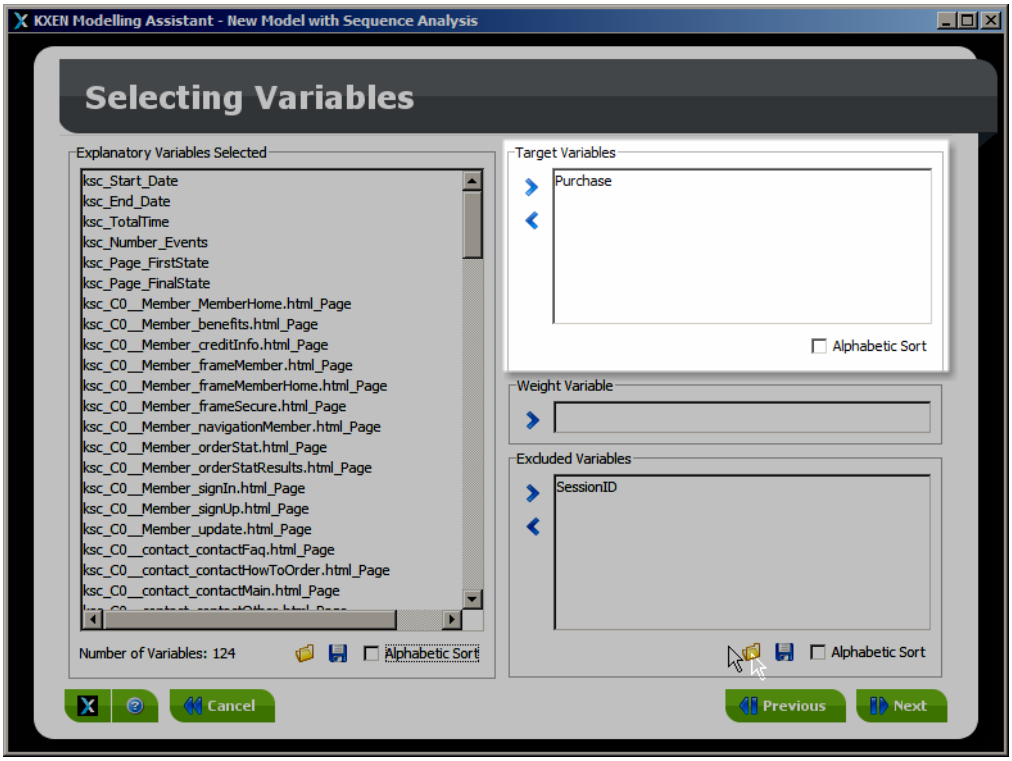
- one or more **Targets Variables**,
- possibly a **Weight Variable**,
- and the **Explanatory Variables**.

For this Scenario

- Keep **Purchase** as the target.
- Use the **session_continue_skip.csv** file to select the variables to exclude. This list of variables includes the information that is not known about a session until a purchase has occurred or is very likely to occur. For this Web site, the checkout process included five order pages. The presence of any of the five order pages in the log indicates that they have already started the checkout process. The presence of order5.tmpl indicates that a purchase has occurred. Since the goal of the analysis is to gain new insights into what behaviors lead to a purchase, these order pages and other similar information must be excluded from the analysis.

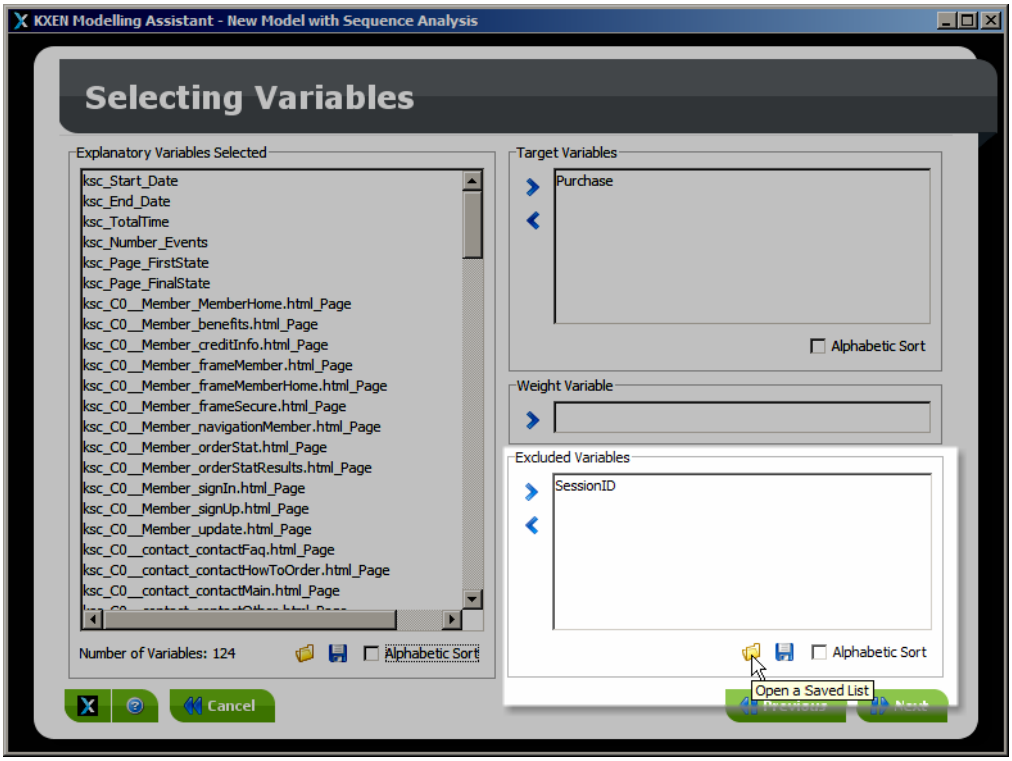
☑ To Select a Target Variable

On the screen **Selecting Variables**, in the section **Explanatory Variables Selected** (left hand side), select the variables you want to use as target variables.

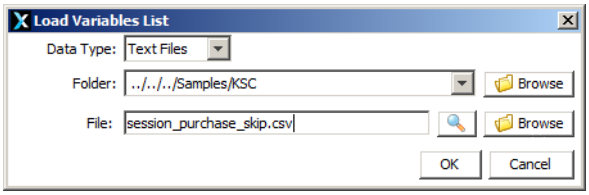


☑ To Exclude Explanatory Variables

1 On the screen **Selecting Variables**, click the button **Open a Saved List** located under the section **Excluded Variables**.



2 The window **Load Excluded Variables List** opens. In the **Variables** field, select the file containing the variables to skip.



3 Click the **OK** button, the window closes. The list of excluded variables has been populated.

Setting the Number of Clusters

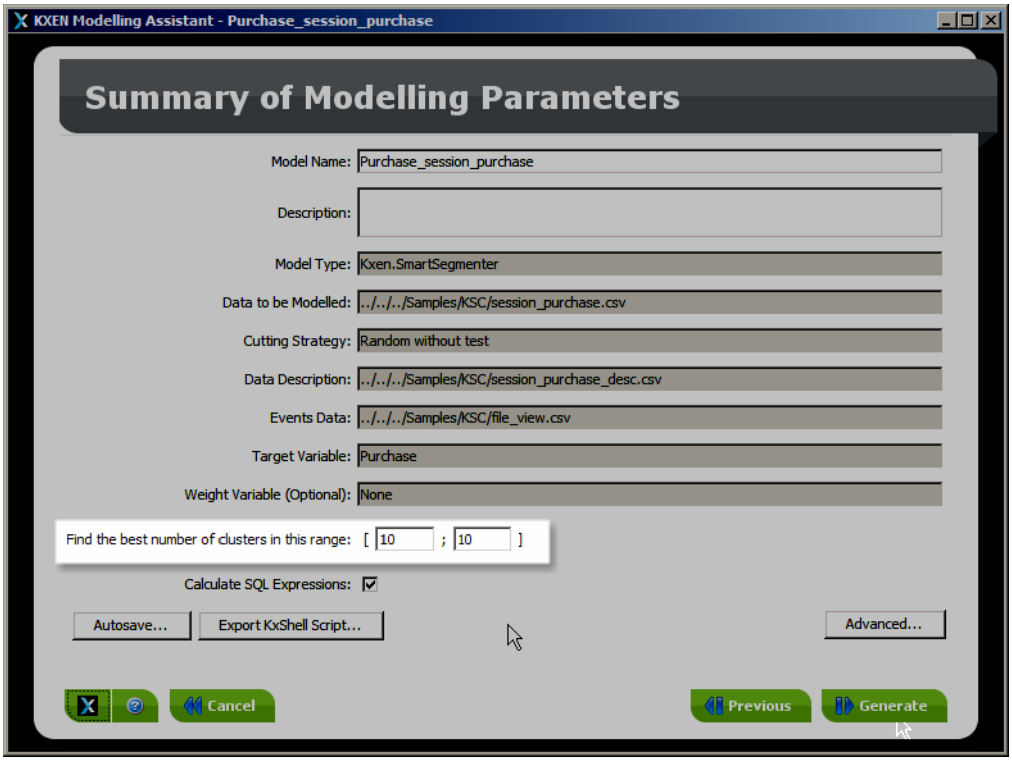
Before generating the model, you need to set the number of clusters you want to create.

For this Scenario

- Set the number of clusters to **10**, which is the default number.

☑ To Set the Numer of Clusters

In the panel **Summary of Modelling Parameters**, type the number of clusters you want to generate in the field **Find the best number of clusters in this range**.



Step 3 - Generating and Validating the Model

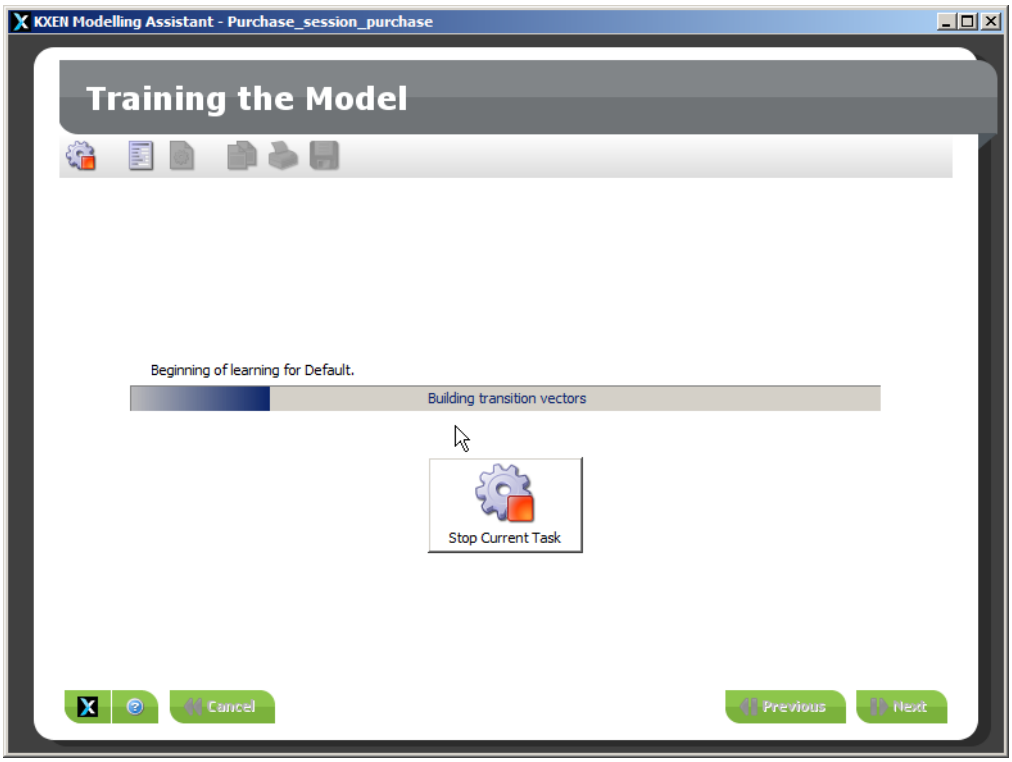
Generating the Model

Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the quality indicator KI and the robustness indicator KR:

- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue.
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your data set and your business issue, and then generate new, more powerful models.

☑ To Generate the Model

- On the screen **Summary of Modelling Parameters**, click the **Generate** button.
  The screen **Training the Model** will appear. The model is being generated. A progress bar will allow you to follow the process.



Validating the Model

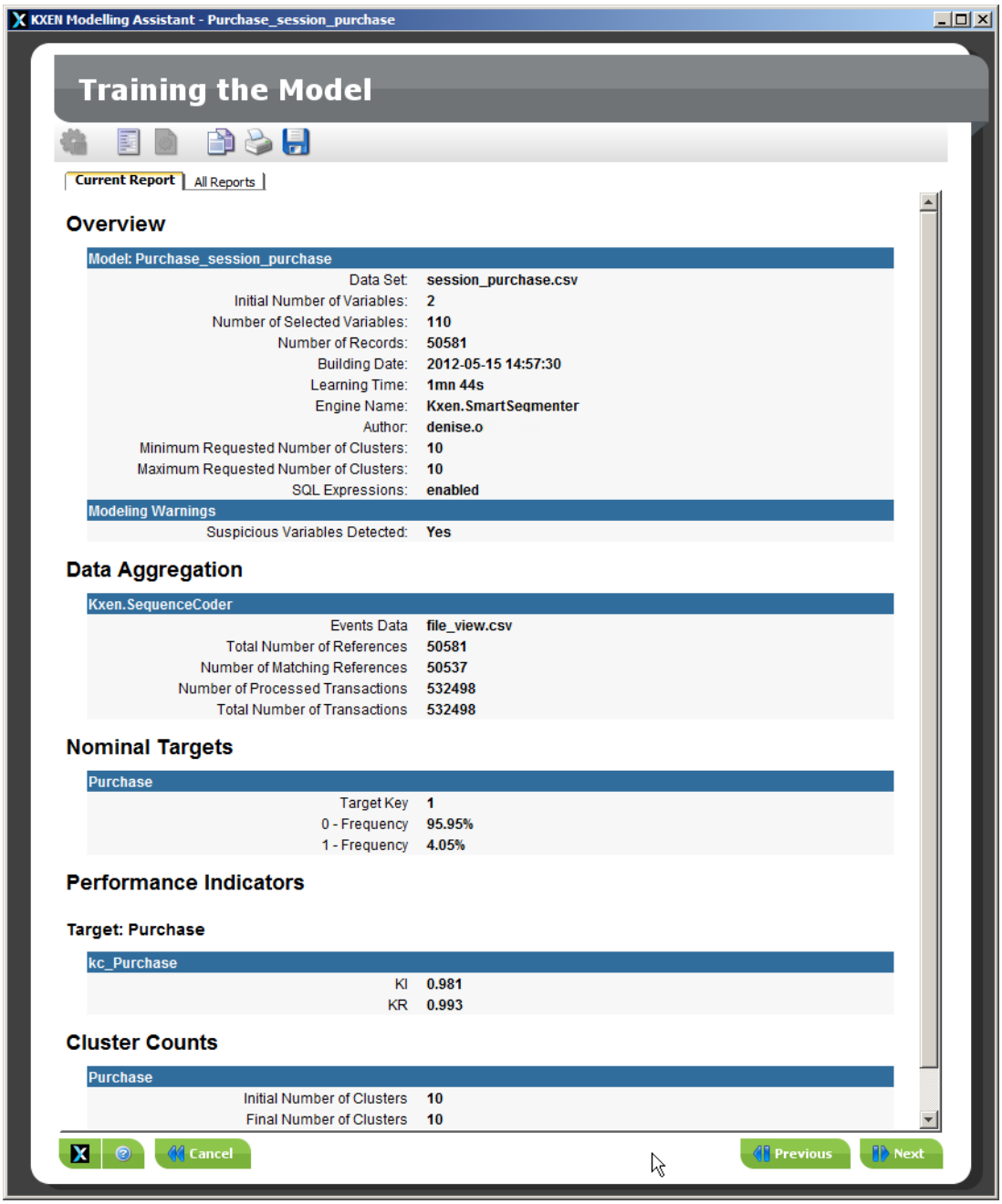
Once the model has been generated, you must verify its validity by examining the performance indicators:

- The quality indicator KI allows you to evaluate the explanatory power of the model, that is, its capacity to explain the target variable when applied to the training data set. A perfect model would possess a KI equal to 1 and a completely random model would possess a KI equal to 0.
- The robustness indicator KR defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new data set. In other words, the degree of robustness corresponds to the predictive power of the model applied to an application data set.

For this Scenario

The model generated possesses:

- A quality indicator **KI** equal to **0.98**,
- A robustness indicator **KR** equal to **0.99**.



This means that Clustering found a reliable grouping (KR is greater than 0.90) that does a reasonable job of partitioning the purchasing visitors and the non-purchasing visitors (KI of 0.98). It is safe to look at the descriptive results of the segmentation to gain insight.

Step 4 - Analyzing and Understanding the Model

Segment Descriptions

On the screen **Cross Statistics**, you can look at the logical definition and/or the cross statistics of each variable to gain an understanding of what kind of visitors belong to each cluster. Three clusters are particularly informative for your business problem, which is to determine which kind of population you should try to attract to increase your profit :

- the two clusters that have the highest conversion rates,
- the cluster that has the lowest conversion rate.

The chart below summarizes these clusters, and gives them each a label based on the cluster definition:

Freq.	Conv.	Definition	Label
1.9%	31.4%	/shop/shipChart.html]0;5]	Shippers
3.5%	25.4%	/welcome.html [1;20]	Members
11.8%	0.1%	/holiday/holidaySweeps.tmpl [1]	Sweepstakers

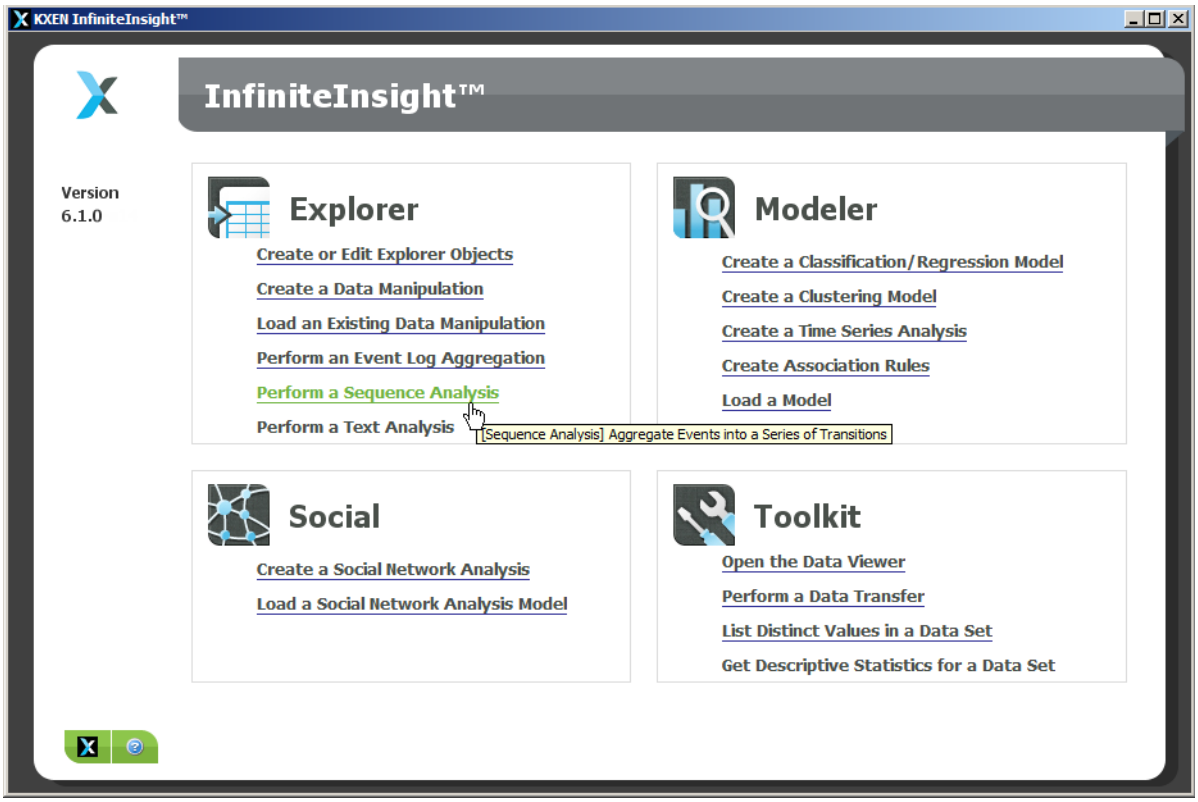
The cluster **Shippers** is defined by sessions in which the shipping chart (**/shop/shipChart.htm**) has been seen between 1 and 5 times. Actually, this cluster does not give you much information. It just tells you that visitors that go to the shipping chart will probably make a purchase, which is rather logical. If you don't intend to buy, why would you look at the shipping information ?

The cluster **Members** is more informative. It shows that people visiting the member home page (**welcome.html**) are more likely to buy. This is an interesting piece of information. It means that members are more likely to make a purchase than other visitors. So increasing the number of members should increase your profit.

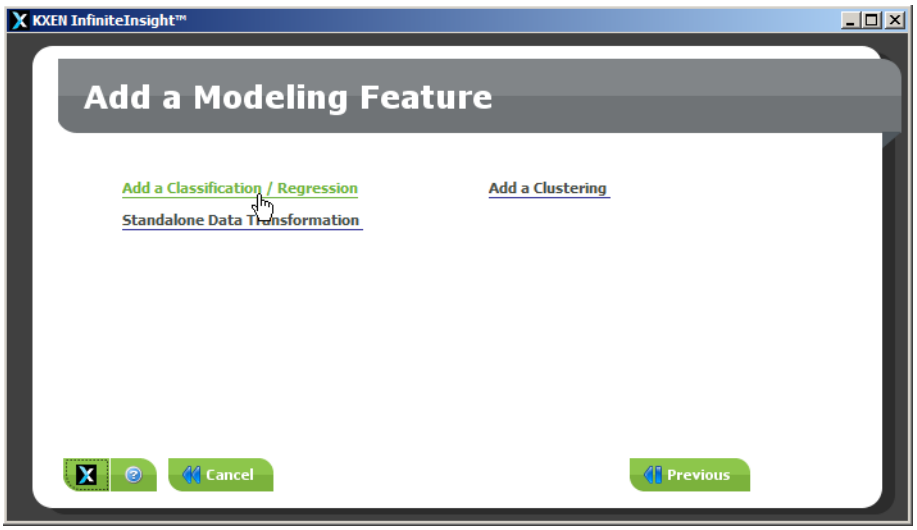
The cluster **Sweepstakers** gives you information on a previous attempt at increasing the number of purchase through a sweepstake. You can see that only 0.1% of the people visiting the sweepstake page actually make a purchase. You can infer from this that your previous campaign had the effect opposite to the one expected.

Scenario 2: Predict End of Session Using Intermediate Sequences

1 In *InfiniteInsight™* main menu, select the option **Perform a Sequence Analysis** in the **Explorer** section.



2 The screen **Add a Modeling Feature** is displayed.



3 Click on the option **Add a Classification / Regression**.

Note - When building a model you can either simply analyze the sequences or add extra transformations such as a Classification/Regression (*InfiniteInsight™ Modeler - Regression/Classification*) or a Clustering/Segmentation (*InfiniteInsight™ Modeler - Segmentation/Clustering*).

IN THIS CHAPTER

Step 1 - Selecting the Data 33
Step 2 - Defining the Modeling Parameters 33
Step 3 - Generating and Validating the Model 34
Step 4 - Analyzing and Understanding the Model 36

Step 1 - Selecting the Data

To know how to select and describe the data go to section **Selecting the Data** (on page 8) and **Describing the Data** (on page 9) in **Scenario 1**.

For this Scenario

- Select the **Random** cutting strategy.
- Use the file **session_purchase.csv** as the reference file and use the file **session_purchase_desc.csv** as its description file.
- Select the file **file_view.csv** and use the description file **file_view_desc.csv**.

Step 2 - Defining the Modeling Parameters

Setting Sequence Coding Parameters

For this Scenario

- Select the **SessionID** column as the join column for both the log and reference data sets.
- Select **Time** as the **Log Date Column**.
- Check the option **Intermediate Sequences**.
- In the advanced parameters, keep **75%** of the hits.

Note - To know how to set the parameters go to section **To Set the Parameters** (on page 16) in scenario 1.

Selecting Sequence Coding Statistics

For this Scenario

In this scenario, you decide to calculate for each session which pages have been visited on the web site and what page led the internaut to another. By adding page transactions count to the model, more information on the internauts behavior will appear.

You decide to calculate for each session which pages have been visited first and last on the web site and what pages had been visited in between. That way, you should be able to determine when a visitor is going to leave your web site and decide on which pages to make a $5 reduction offer to keep the visitor and encourage him to make a purchase.

You must use the following settings.

For the variable **Page**, select the function **FirstLast**, which will create two states columns for each session, one containing the first page visited, the other the last page visited.

Note - To know more about Sequence Coding Statistics, go to section **Selecting** *InfiniteInsight™ Explorer - Sequence Coding* **Statistics** (see "**Selecting Sequence Coding Statistics**" on page 22) in scenario 1.

Checking the Transactions

For this Scenario

After the transactions are checked, Sequence Coding should have kept **98** state columns for the **Page** variable.

Selecting Variables

For this Scenario

- Use the **session_continue_skip.csv** file to select the variables to exclude.
- Use **KSC_Session_continue** as the target and remove **Purchase** from the targets.

Note - To know how to select variables, go to section **Selecting Variables** (see "**For this Scenario**" on page 26) in scenario 1.

Step 3 - Generating and Validating the Model

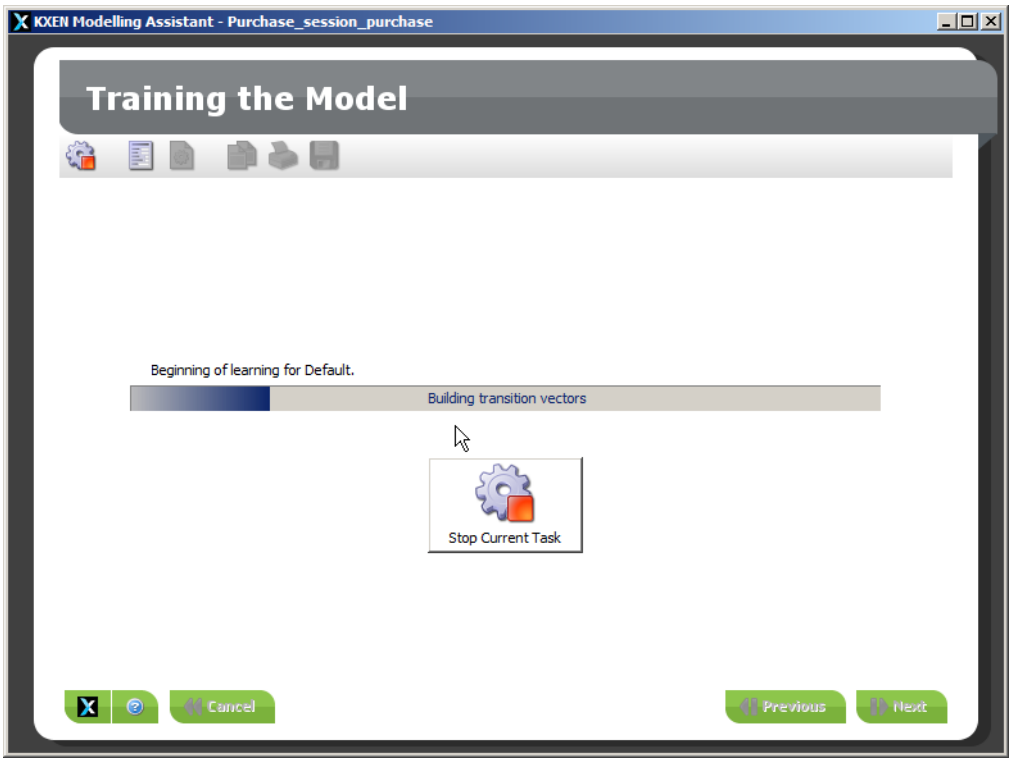
Generating the Model

Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the quality indicator KI and the robustness indicator KR:

- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue.
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your data set and your business issue, and then generate new, more powerful models.

☑ To Generate the Model

- On the screen **Summary of Modelling Parameters**, click the **Generate** button.
  The screen **Training the Model** will appear. The model is being generated. A progress bar will allow you to follow the process.



Validating the Model

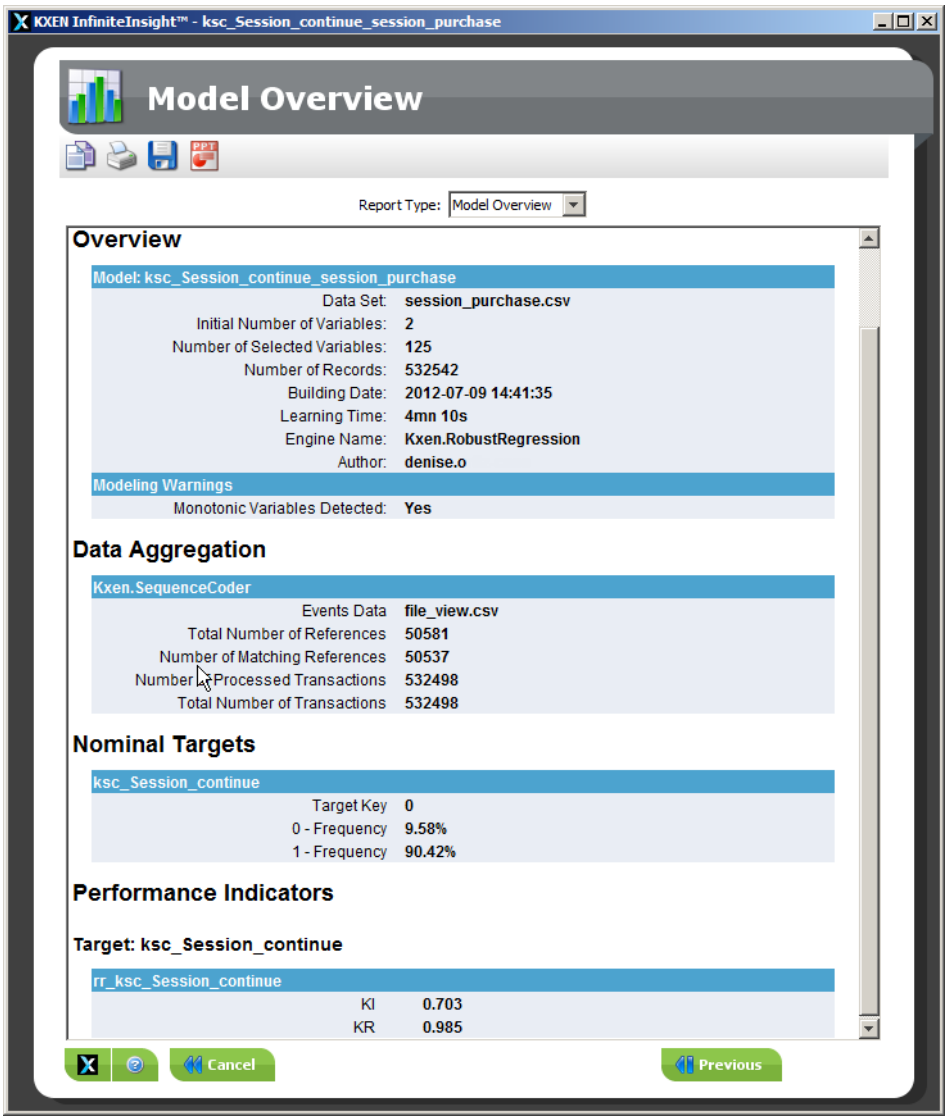
Once the model has been generated, you must verify its validity by examining the performance indicators:

- The quality indicator KI allows you to evaluate the explanatory power of the model, that is, its capacity to explain the target variable when applied to the training data set. A perfect model would possess a KI equal to 1 and a completely random model would possess a KI equal to 0.
- The robustness indicator KR defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new data set. In other words, the degree of robustness corresponds to the predictive power of the model applied to an application data set.

For this Scenario

The model generated possesses:

- A quality indicator **KI** equal to **0.70,**
- A robustness indicator **KR** equal to **0.98**.

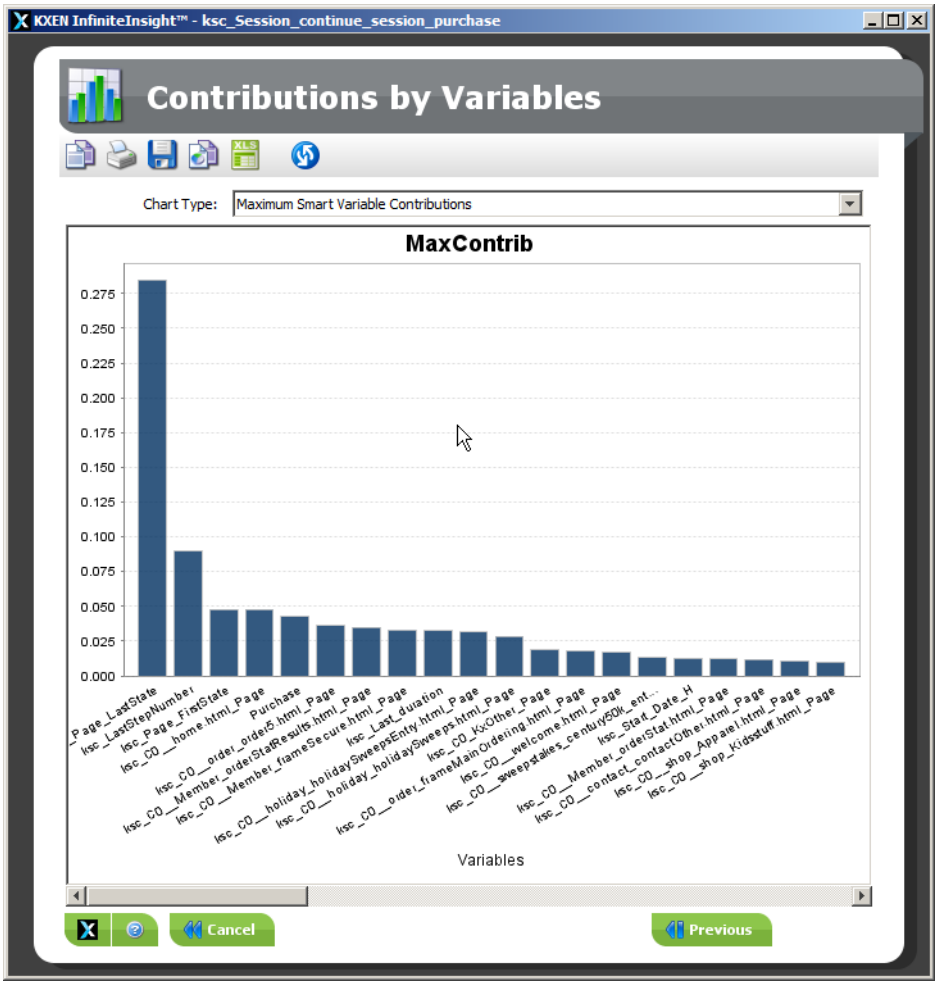


This means that Classification/Regression found a robust model (KR is greater than 0.90) that does a reasonable job of predicting the end of a session (KI of 0.70). It is safe to look at the variables contributions to gain insight.

Step 4 - Analyzing and Understanding the Model

Contributions by Variables

The following graph presents the variables contributions.



The pages having the more impact (positive or negative) on the buying act are listed in the following table.

Page viewed	This variable indicates...
KSC_Page_LastState	the last page the internaut has viewed before ending his session
KSC_Last_duration	duration of the session from the first page viewed to the previous state
KSC_LastStepNumber	the number of pages the internaut has viewed before ending his session
Count_holidaySweepsEntry.html	the number of time the page holidaySweepsEntry (access to holiday promotions) has been viewed

The impact of each page on the purchase is detailed in section **Significance of Categories**.

Significance of Categories

KSC_Page_LastState

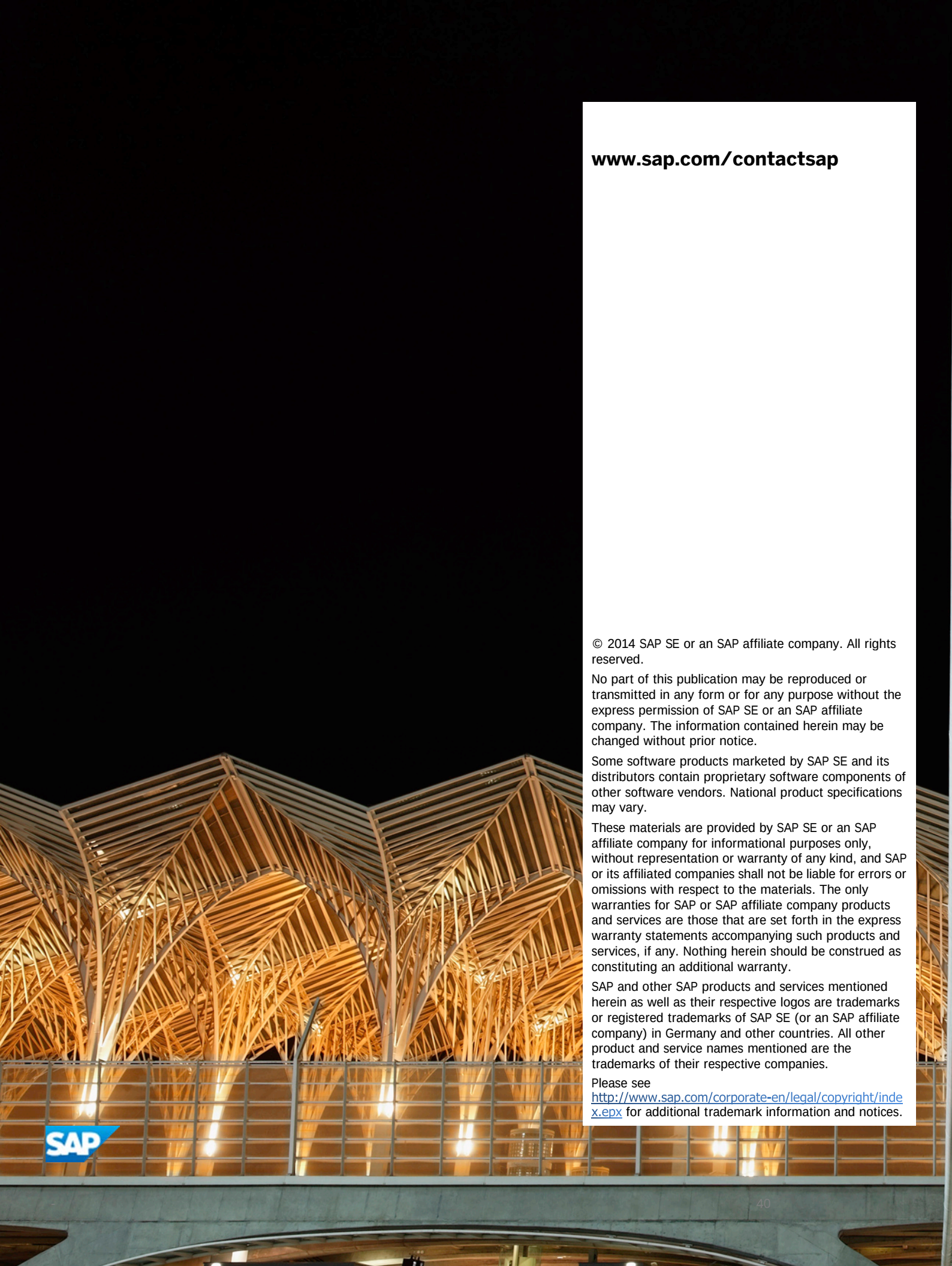
This is by far the strongest predictor. This is similar to a low order Hidden Markov Model, where the current state is used to predict the next one.

Last_duration and **LastStepNumber**

The length of the session and the number of pages viewed are also important. If the internaut has viewed only one page, he has not yet entered the site and may end his session because the site may not seem of interest to him, but if he has viewed more than 12 pages, he has probably found what he was looking for and will end his session. If he has seen between 2 and 11 pages, he is probably shopping and thus should continue his session.

Count_holidaySweepsEntry.html

If the page has been viewed it is a good indicator that the session will continue. Since this page is the entry point of a holiday promotion, the internaut will at least go to the promotion page.



www.sap.com/contactsap

© 2014 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company. The information contained herein may be changed without prior notice.

Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.

Please see http://www.sap.com/corporate-en/legal/copyright/index.epx for additional trademark information and notices.